Лабораторная работа №3

Тема: Анализ качества однородных подмножеств данных.

Цель работы: получить навыки проверки качества разбития исходной совокупности данных на однородные подмножества.

Методические указания к выполнению работы

Разбитие на полностью отделенные подмножества характеризуется тем, что объекты внутри подмножества слабо различаются между собой (т.е. сильно коррелируемые), а объекты разных подмножеств имеют значительные отличия. Разбитие на частично отделенные подмножества характеризуется тем, что объекты разных подмножеств не имеют значительных отличий между собой. Степень отличия можно определить с помощью диаметров подмножеств и расстояний между подмножествами.

Диаметр подмножества P_i определяется по формуле:

$$d_{ii} = \max_{n_i, n_j \in P_i} \left\{ d\left(n_i, n_j\right) \right\}, \tag{3.1}$$

Расстояние между подмножествами P_i и P_j определяется по формуле:

$$d_{ij} = \min_{n_i \in P_i, n_j \in P_j} \left\{ d\left(n_i, n_j\right) \right\}. \tag{3.2}$$

В результате получаем матрицу характеристик разбития совокупности на однородные подмножества: $D = \left(d_{ij}\right)_{n \times n}$.

Дальше определяется предельное значение $d^* = \max_i \left\{ d_{ii} \right\}$. Значения матрицы, превышающие d^*, к сведению не принимаются, поскольку характеризуют сильно отделенные подмножества. То есть отличия между разными элементами этих подмножеств больше, чем отличия внутри подмножеств. Для удобства дальнейшей обработки данных матрица D превращается в матрицу D' в соответствии с указанным выше условием:

$$d'_{ij} = \begin{cases} d_{ij}, & d_{ij} \le d^*, \\ 0, & d_{ij} > d^*. \end{cases} \tag{3.3}$$

При разбитии объектов на непустые пересекающиеся множества, не всегда получается результат с хорошим, с точки зрения отделения подмножеств, разбитием. Особенности выделенных подмножеств можно описать с помощью внутренних и внешних коэффициентов подобия, которые вычисляются на основе матрицы *D* расстояний между объектами.

Степень внутреннего подобия подмножества P_i определяется по формуле:

$$a'_{ij} = \begin{cases} d^* - \dfrac{1}{k_i} \sum\limits_{n_i, n_j \in P} d\left(n_i, n_j\right), & k_i > 0, \\ d^*, & k_i = 0, \end{cases} \tag{3.4}$$

где k_i – количество пар элементов подмножества P_i. Второе выражение формулы (3.4) касается одноэлементных подмножеств.

Для измерения степени подобия разных подмножеств используется величина:

$$a'_{ij} = d^* - \frac{1}{k_{ij}} \sum_{n_i \in P_i, n_j \in P_j} d^*\left(n_i, n_j\right), \tag{3.5}$$

где $d^*\left(n_i, n_j\right) = \begin{cases} d\left(n_i, n_j\right), & d\left(n_i, n_j\right) \le d^*, \\ d^*, & d\left(n_i, n_j\right) > d^*; \end{cases}$ k_{ij} – количество пар элементов подмножеств P_i и P_j, для которых выполняется условие $d\left(n_i, n_j\right) \le d^*$. Высокие значения показателей свидетельствуют о подобии между объектами как внутри подмножества (для a'_{ii}) так и между подмножествами (для a'_{ij}).

В практическом применении такие коэффициенты неудобны, потому их превращают таким образом:

$$a_{ij} = \frac{a'_{ij}}{d^*}, \tag{3.6}$$

Очевидно, что $0 \le a_{ij} \le 1$. Чем более близкое значение коэффициента к единице, тем высшая степень подобия между объектами подмножеств.

По значениям матрицы A можно построить показатель, измеряющий степень когерентности (сцепление) элементов разных подмножеств. Коэффициент простой когерентности имеет вид:

$$k_{ij} = \frac{a_{ij}}{a_{ii} + a_{jj}}. \tag{3.7}$$

Чем большая величина k_{ij}, тем меньше отличаются между собой подмножества. Дополнение этой величины к единице даст другую величину – коэффициент отделения:

$$s_{ij} = 1 - k_{ij}. \tag{3.8}$$

Также интересен ответ на вопрос, в какой степени каждая из подмножеств не похожая на другие подмножества той же совокупности. Для этого служит коэффициент индивидуальной когерентности:

$$k_{ii} = \sum_{i \ne j} k_{ij}. \tag{3.9}$$

Чем большие значения величины k_{ii}, тем большее сходство данного подмножества с другими. К полностью отделенным подмножествам относят те, для которых $k_{ii} = 0$. В противные случаи подмножества называют частично отделенными.

Рассчитаны по формулам (3.6) и (3.8) коэффициенты образуют матрицу когерентности K. Она предоставляет возможность оценить качество разбития исходной совокупности данных на подмножества. Это делается с помощью обобщенного коэффициента когерентности:

$$k_{ii} = \frac{1}{s} \sum_{j>i} \sum_{i} k_{ij}. \tag{3.10}$$

где s – количество пар подмножеств, для которых $k_{ij} > 0$.

Сумма берется для элементов матрицы *K*, расположенных выше главной диагонали. По существу, обобщенный коэффициент когерентности являет собой среднее значение коэффициентов простой когерентности. Он характеризует общий уровень сближения (или отдаление) всех подмножеств данной совокупности. Чем большее значение коэффициента, тем высший уровень сближения подмножеств.

Задание для самостоятельного выполнения

Для полученного в ходе выполнения лабораторной работы №1 разбитие данных на однородные подмножества оценить качество их отделения. Для каждого типа разбития (изотропного, изотонического, изоморфического) рассчитать:

1) значение матрицы *D*;

2) значения матрицы *A*;

3) коэффициенты когерентности (простые, индивидуальные и обобщенные).

По результатам расчетов сделать выводы по поводу качества разбития.

Рекомендуемая литература

1. Гмурман В.Е. Теория вероятностей и математическая статистика. Уч. пособие для втузов. – М.: Высш. школа, 2002. – 479 с.

2. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. Уч. пособие для втузов. – М.: Высш. Школа, 2002. –ы 400 с.

3. Жалдак М.И. Теория вероятностей с элементами информатики. Практикум: Уч. Пособие / М.И. Жалдак, А.Н. Квитко / Под общ. ред. М.И. Ядренко. – К.: Вища шк., 1989. – 263 с.

4. Ивченко Г.И. Математическая статистика: учеб. пособие для втузов / Г.И. Ивченко, Ю.И. Медведев. – М.: Высш. шк., 1984. – 248 с.

5. Методичні рекомендації до виконання лабораторних робіт з навчальної дисципліни "Моделювання систем" для студентів напрямку підготовки 0804 "Комп’ютерні науки" всіх форм навчання / укл. В. М. Задачин, І. Г. Конюшенко. – Харків : Вид. ХНЕУ, 2007. – 96 с.

6. Тарасова П.В. Введение в математическое моделирование: учеб. пособие для вузов / под ред. П.В. Тарасова. – М.: Интермет Инжиниринг, 2000. – 200 с.